SPECTRa-T :

Semantic Web Data Repositories from Chemistry e-Thesis Data Mining

Jim Downing[1], Peter Murray-Rust[1], Diana Stewart[1], Alan Tonge[1], Joe Townsend[1], Peter Morgan[2], Henry S.Rzepa[3] & Matt J.Harvey[4]

1. Unilever Centre for Molecular Informatics, Department of Chemistry, Lensfield Rd., Cambridge CB2 1EW, U.K.
2. Cambridge University Library, West Rd., Cambridge CB3 9DR, U.K
3. Department of Chemistry, Imperial College London, Exhibition Rd., London SW7 2AY, U.K
4. High Performance Computing Unit, ICT, Imperial College London, Exhibition Rd., London SW7 2AZ, U.K

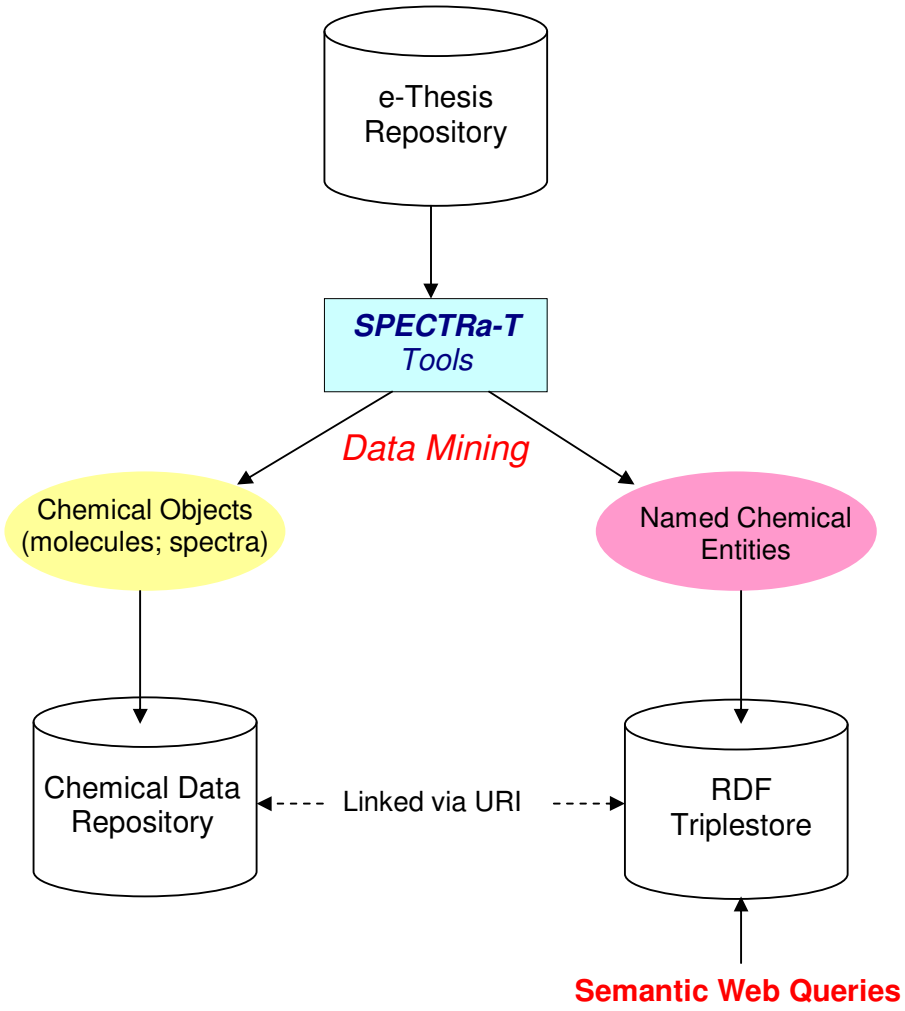
There is a major effort in many countries to provide Open Access repositories of academic theses. For example, the SURF Foundation[1] in the Netherlands and the DART Europe portal[2] provide access to many thousands of e-theses. Although some of these theses are 'born digital', many are not, and yet there are great expectations to realise access to and re-use of the extensive unpublished data and metadata contained within them. However, for domain-specific metadata, large-scale human annotation is impossibly expensive. There are additional problems associated in dealing with OCR-scanned images of text and the non-linear storage of text within PDF documents[3], the *de facto* format for repository deposition. This project has therefore developed text mining tools that address the need to extract the wealth of experimental data currently untapped in scientific theses, focussing on chemistry research data in molecular and related subjects.

Much of the data (synthetic, spectral, computational and even crystallographic) generated by postgraduate researchers in chemistry and related departments are conventionally reported in theses. Although such theses might describe up to 50 novel chemical syntheses, much of this is not communicated in peer-reviewed publication to the scientific community in an appropriate form (numbers are reduced to points on diagrams, tables are converted to graphs in pixel form) and a significant proportion (anecdotally estimated at 50%) is never formally submitted at all. Although the bare essentials of the synthesis are published, the detailed experimental recipes (as found in the thesis) are often omitted. As we have found in the recently-completed SPECTRa project[4,5], researcher compliance with deposition of original primary data is a major obstacle. We have noted that the social aspects (ownership, fear of premature publication, etc.) were probably more important than the technical ones (e.g. lack of software) and militated against rapid deposition or high-compliance. Theses, however, have few of these social constraints: a student must comply with regulations, must provide all supporting information to examiners if required, must assemble the data to a given quality metric.

The SPECTRa-T project[6] has explored the potential for text mining e-theses in PDF and Office Open XML ('docx') formats using OSCAR3[7] to capture sufficient data and metadata as RDF[8] and URI-linked CML[9] chemical objects, enabling:

- routine and automatic extraction of Chemical Objects (*e.g.* molecules, spectra) and named chemical entities in high volumes, transformation into metadata and their capture into data repositories and triplestores.
- exploration of the viability of RDF-based semantic querying.
- review of current document format practice in the deposition of chemistry theses and how this influences ease of data extraction

We have succeeded in extracting data from chemistry e-theses and re-using selected terms as machine-readable metadata. APP-enabled[10] repositories are potentially an effective means of capturing, preserving, and disseminating the associated molecular data in accordance with Open Access principles.



Overview of SPECTRa-T data mining architecture

The SPECTRa-T suite will include a facility to conduct semantic searches, which differ from free text searches in that any of the three components of an RDF data triple can be specified. Such searches allow "chaining", so that the results of the first search can be piped into a second, and so on. The level of sophistication of such searches far exceeds that of normal free text search.

Acknowledgements

We thank the Joint Information Systems Committee (JISC) for financial support.

References

1. *http://www.surf.nl/en/home/index.php*
2. *http://www.dartington.ac.uk/dart/*
3. Portable Document Format. PDF/A has ISO approval as a standard for long-term preservation of electronic documents.
4. *http://www.lib.cam.ac.uk/spectra/Questionnaire.html*
5. *http://www.lib.cam.ac.uk/spectra/FinalReport.html*
6. *http://www.lib.cam.ac.uk/spectra-t/*
7. Corbett P., Murray-Rust P., *High-Throughput Identification of Chemistry in Life Science Texts*, Computational Life Sciences II, pp107-118 (Springer, Berlin, 2006)
8. Resource Description Framework: RDF Primer, *http://www.w3.org/TR/REC-rdf-syntax*
9. Murray-Rust P., Rzepa H.S. and Wright M., *Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content*, New J. Chem., **2001**, 618-634.
10. SWORD APP Profile, *http://www.ukoln.ac.uk/repositories/digirep/index/SWORD_APP_Profile_0.4*